

ORIGINAL ARTICLE OPEN ACCESS

Real-World Evidence BRIDGE: A Tool to Connect Protocol With Code Programming

Albert Cid Royo  | Roel Elbers JHJ | Daniel Weibel | Vjola Hoxhaj | Zeynep Kurkcuoglu | Miriam C. J. Sturkenboom | Tiago A. Vaz | Constanza L. Andaur Navarro 

Department of Data Science and Biostatistics, Julius Center for Health Science and Primary Care, University Medical Center of Utrecht, Utrecht University, Utrecht, The Netherlands

Correspondence: Albert Cid Royo (a.cidroyo@umcutrecht.nl)

Received: 30 July 2024 | **Revised:** 18 October 2024 | **Accepted:** 7 November 2024

Funding: The authors received no specific funding for this work.

Keywords: drug safety | drug utilization | electronic health records | pharmacoepidemiology | vaccines

ABSTRACT

Objective: To enhance documentation on programming decisions in Real World Evidence (RWE) studies.

Materials and Methods: We analyzed several statistical analysis plans (SAP) within the Vaccine Monitoring Collaboration for Europe (VAC4EU) to identify study design sections and specifications for programming RWE studies. We designed a machine-readable metadata schema containing study sections, codelists, and time anchoring definitions specified in the SAPs with adaptability and user-friendliness.

Results: We developed the RWE-BRIDGE, a metadata schema in form of relational database divided into four study design sections with 12 tables: Study Variable Definition (two tables), Cohort Definition (two tables), Post-Exposure Outcome Analysis (one table), and Data Retrieval (seven tables). We provide a guide to populate this metadata schema and a Shiny app that checks the tables. RWE-BRIDGE is available on GitHub (github.com/UMC-Utrecht-RWE/RWE-BRIDGE).

Discussion: The RWE-BRIDGE has been designed to support the translation of study design sections from statistical analysis plans into analytical pipelines and to adhere to the FAIR principles, facilitating collaboration and transparency between researcher and programmers. This metadata schema strategy is flexible as it can support different common data models and programming languages, and it is adaptable to the specific needs of each SAP by adding further tables or fields, if necessary. Modified versions of the RWE-BRIDGE have been applied in several RWE studies within VAC4EU.

Conclusion: RWE-BRIDGE offers a systematic approach to detailing variables, time anchoring, and algorithms for RWE studies. This metadata schema facilitates communication between researcher and programmers.

1 | Introduction

In pharmacoepidemiology, retrospective studies on available Real-World data (RWD) are common. This type of Real-World Evidence (RWE) focuses on drug utilization, surveillance, and drug and vaccine safety and effectiveness. RWE relies on the rapid analysis of RWD, allowing for timely insights into treatment patterns, adverse events, and vaccine performance in

diverse populations. The current practice is to work collaboratively with several databases using distributed analysis across Europe, America, and Asia [1–5].

The increased generation of RWE has led to a demand for standardization and transparency in analytical pipelines. Initiatives like RECORD-PE [6], START-RWE [7], SPACE framework [8], and the HARPER [9] advocate for more transparent and clear

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial](https://creativecommons.org/licenses/by-nc/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

© 2024 The Author(s). *Pharmacoepidemiology and Drug Safety* published by John Wiley & Sons Ltd.

Summary

- RWE-BRIDGE addresses the methodological gap in translating study protocols into executable analytical pipelines for real-world evidence (RWE) studies increasing transparency and communication between researchers and scientific programmers. Moreover, this methodology aligns with HARmonized Protocol Template to Enhance Reproducibility (HARPER) and Structured Preapproval and Postapproval.
- Comparative study design framework to generate valid and transparent real-world Evidence (SPACE) protocols.
- RWE-BRIDGE provides a ready-to-use, machine-readable, and functional solution that streamlines the development of RWE analytical pipelines. It consists of 12 different tables divided into four sections: Study Variable Definition (two tables), Cohort Definition (two tables), Post-Exposure Outcome Analysis (one table), and Data Retrieval (seven tables).
- RWE-BRIDGE is compatible with any programming language and can support multiple Common Data Models, enabling wide applicability when using federated analysis.
- Our tool adheres to FAIR principles by enhancing the findability, accessibility, interoperability, and reusability of RWE resources. RWE-BRIDGE is available on GitHub (github.com/UMC-Utrecht-RWE/RWE-BRIDGE).

documentation and reporting of RWE studies. However, examples of unreproducible studies are available in the scientific literature, leading to skepticism about the quality of RWE [10]. While statistical analysis plans (SAPs) provide detailed information on study design and analysis elements, they often need clearer, plain-language descriptions for creating study variables, time anchoring, algorithms, risk windows, lookback periods, codelists, and rules. Consequently, programmers require continuous research support to make scripting decisions while building analytical pipelines.

To enhance the documentation and translation of SAPs into code, we need a tool that succinctly encapsulates the decisions made during the analytical pipeline's programming. Despite existing literature on best practices for executing and reporting pharmacoepidemiological studies using RWD, there is a lack of tangible methods or tools for converting the SAP information into a machine-readable format for programming analytical pipelines [6, 7, 11–14].

Following the recommendations from RECORD-PE [6] and REPEAT [15] initiatives, we developed a metadata schema that (1) provides an adaptable and transparent solution to accommodate study specifications from the SAP; (2) can automatically incorporate study specifications into analytical scripts; (3) facilitates communication between collaborators in RWE studies; and (4) adheres to the findable, accessible, interoperable, and reproducible (FAIR) principles [16]. In this article, we present the RWE-BRIDGE (Bring Intelligence about Data to Generation

of Evidence) to improve and facilitate the documentation of programming decisions.

2 | Materials and Methods

To design a metadata schema that transforms a SAP into a machine-readable document, we reviewed multiple SAP developed within VAC4EU (Vaccine Monitoring Collaboration for Europe) [17], which generates RWE on vaccines in Europe. RWE generated within VAC4EU uses the ConcePTION common data model (CDM), which is a CDM requiring only syntactic harmonization. In contrast, semantic harmonization is conducted as part of the study's analytical script. Further details are described elsewhere [18].

Studies within VAC4EU apply an analytical pipeline wherein researcher and statisticians write the protocol and SAP. Participating sites (e.g., research partners and data access providers (DAP)) review these documents. DAPs will extract, transform (step T1), and load their local data into a CDM (D2). Programmers are responsible for scripting steps T2 (creation of study variables and study population) and T3 (application of the design). Furthermore, statisticians are responsible for developing the scripts for the estimands (T4), which are later polled together and post-processed (T5) in tables and graphics for study reports (D6); see Figure 1.

Across SAPs, we identified four main relevant study design elements: (1) operationalization of study variable (e.g., codes and algorithms) for outcomes, exposure, covariates, and inclusion/exclusion criteria; (2) definition of the population/cohort and time anchoring (e.g., observation period, lookback period, follow-up window, censoring); (3) data retrieval; and (4) data analysis. We expected each main study design element to comprise of one or more metadata tables, each consisting of various fields. To determine the necessary tables and fields, we formulated a set of questions based on generic vaccines Post-Authorisation Safety Studies (PASS). See “table questions” in the [Supporting Information](#) for a correlation between these questions and the proposed metadata schema.

In the SAP, precise identification of population specifications, medical concepts/phenotypes, and algorithms is crucial. Each study variable should be defined in the metadata schema using a codelist, phenotype, or algorithm, its role (i.e., exposure, outcome, covariate), and time anchoring. For instance, when determining a covariate's lookback period, a statement like “A person is considered immunocompromised when X code is found before the index date” could be ambiguous because it lacks a fixed beginning of look-back period that could help determining an assessment period (i.e., recent vs. first). It is necessary to clearly define the lookback period's start and end with the index date as an anchor. Figure 2 shows two machine-readable metadata files logging window information to identify study variables.

Relational databases are powerful and adaptable to various data types, making them the most suitable choice for our metadata schema [19]. A relational database organizes data in tables linked through common fields (foreign keys) for efficient querying and validation (i.e., data integrity). A relational database

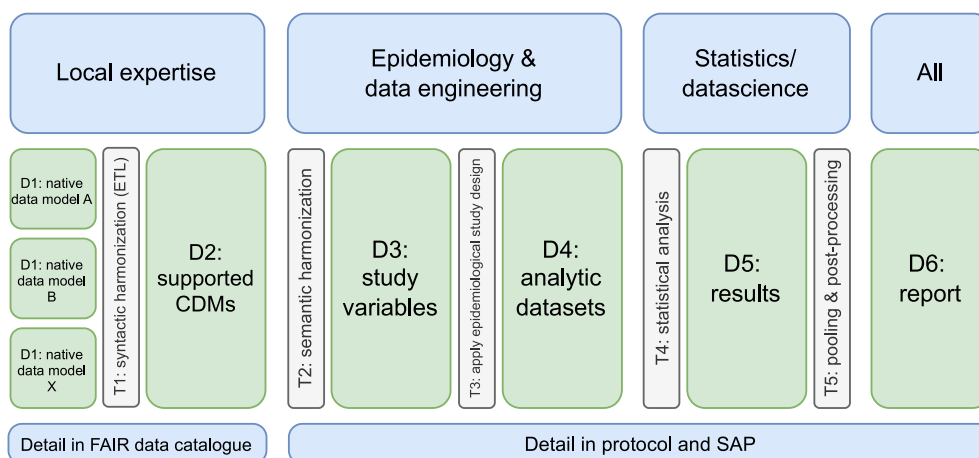


FIGURE 1 | The VAC4EU pipeline is a one-way pipeline that shows all the datasets (D) and data transformation steps (T) applied during a VAC4EU project. The steps are applied consecutively from left to right. The different expertise required in each section of the pipeline can be seen at the top, while the information required for each step and dataset is shown at the bottom. Details on the native data models are necessary at the beginning of the pipeline. These native data models are later ETL'ed. (extract, transfer, and load) into the CDM. Details on the SAP are needed for step T2 to T5.

with a star schema connects the central table to most secondary tables using foreign keys. In this context, a foreign key is a field in a table linked to another table's primary key.

3 | Results

We provide a comprehensive description of the RWE-BRIDGE metadata schema. For the complete structure, see Figure 3. Key terms and definitions can be found in the glossary table in the [Supporting Information](#), which includes a step-by-step guide for manually populating the RWE-BRIDGE.

3.1 | System Description and Structure

The RWE-BRIDGE structure is composed of (1) variable definition with two tables, (2) data retrieval and pre-processing with seven tables, (3) cohort definition with two tables, and (4) data analysis with one table. These four sections interact through a relational metadata schema; see Figure 3, where all tables are interconnected through foreign keys: "Variable ID", "Concept ID", "Cohort name", "DAP name", and "Coding system."

3.1.1 | Variable Definition

This section consists of two tables: Study variables and composite study variables.

3.1.1.1 | Study Variables. This is the core table of the RWE-BRIDGE as it assigns unique IDs to each study variable described in the SAP. The "Variable ID" field serves as a unique ID and foreign key to connect with the Composite Study Variables, Study Variables Ranges, and the Study cohort tables. The concept field connects with the Codelists, DAP Specific Concept Map, and Concept pre-processing tables. A medical concept/phenotype is a list of diagnosis codes and/or integer values

and/or categorical values that conceptually define a study variable. A concept will constitute a study variable when anchored to a reference date and a period of interest defined in the Study Variables table, with the start and end of the lookback period. Exceptions are exposure variables, where the lookback period isn't specified, and composite variables, where the sub-study variables determine the lookback period. The Study Variables table includes fields for specifying variables' roles and time windows. In the Study Variables table, one can specify the programming data type (e.g., boolean, categorical, numerical) and which function is used to select records, for instance, the earliest or latest date in the anchored window period or a count or summation of all records—resulting in a numerical study variable—within the anchored window. In addition, the table includes fields that indicate the variable's role within the study: Exposure, outcomes, or covariates. Finally, each variable can be described in plain language in the field variable description, providing further details to "Variable ID."

3.1.1.2 | Composite Study Variables. This metadata table defines an extra level of combination for study variables by using simple logical formulas to combine them into a composite variable. The Composite Study Variables table only allows combining study variables with OR, AND, and NOT logic. Before creating a composite study variable, the concepts for the study variables should have been anchored, constituting a study variable per se. For example, the study variable Diabetes Type 1 or 2 can be expressed using a logic formula of having either Diabetes Type 1 OR Diabetes Type 2 OR using diabetes medication. The RWE-BRIDGE allows the medical concepts of a composite variable to have different lookback periods. In the Composite Study Variables table, the AND and AND NOT logics are only possible when the concepts do not have a condition between each other (i.e., an internal anchoring). An example of internal anchoring is a study variable that categorizes whether a person was hospitalized after a COVID-19 infection or not. To ascertain such definition, we need to identify all COVID-19 infection records with a code for hospitalization within a window after

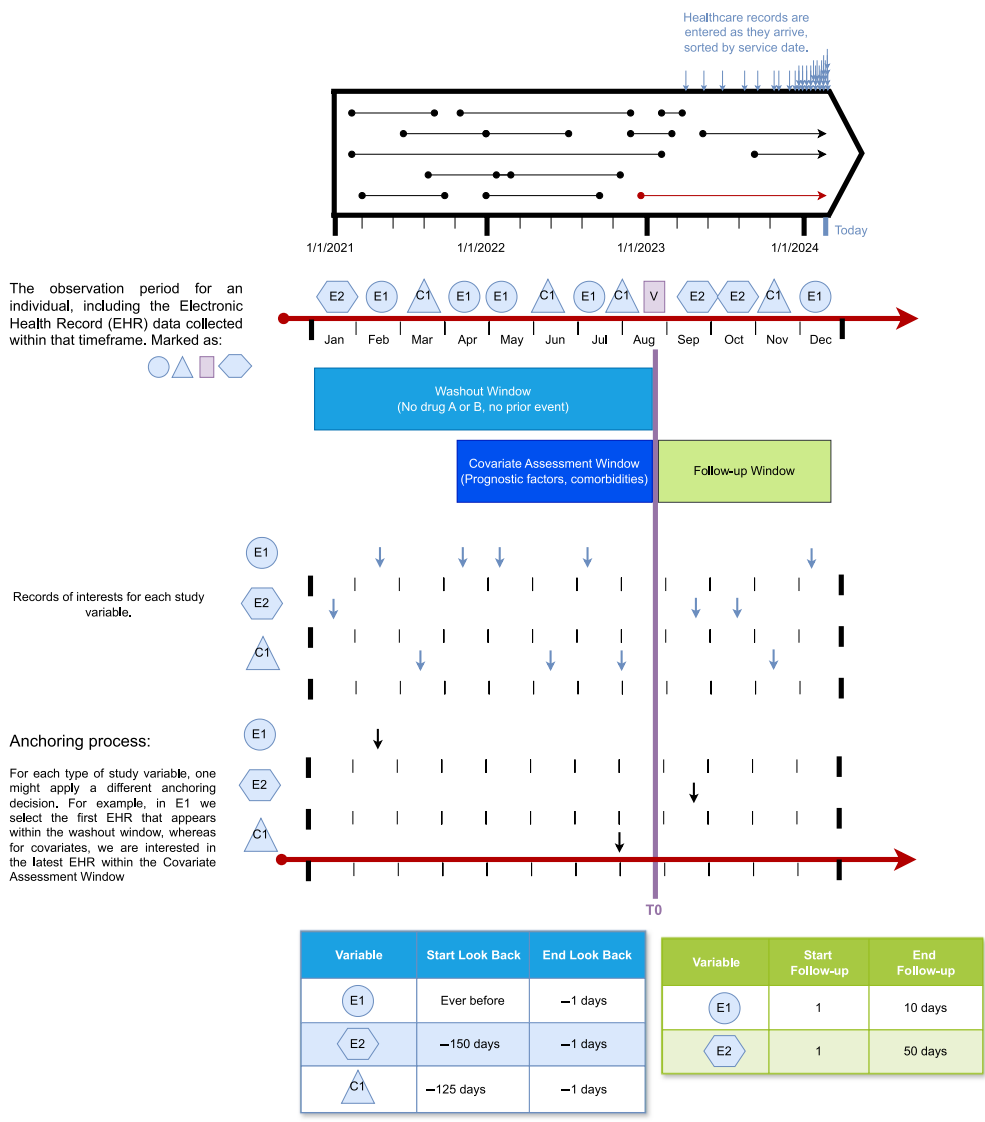


FIGURE 2 | Figure adapted from [12] and expanded with the representation of study variable records and time anchoring process. In the first section of this figure, depicted as horizontal lines, we can see all the observation periods belonging to a person. During observation periods of a person, records are logged into the database. Records are categorized into study variables and then selected for analytical analysis. The second step is the anchoring process, where T0 is shown in purple, and we set the anchoring window. After the second step, we select a date for the study variable after excluding the records that do not fall within the window. We show two options here: Select the earliest or latest records in their respective market, shown in blue. At the bottom, we show how we can define the different characteristics of the time window in a machine-readable format.

their COVID-19 infection record date (anchor within the study variable). In this case, we must add such a definition as a new concept in the Concept pre-processing table.

3.1.2 | Data Retrieval and Pre-Processing

The data retrieval and processing section consists of seven tables.

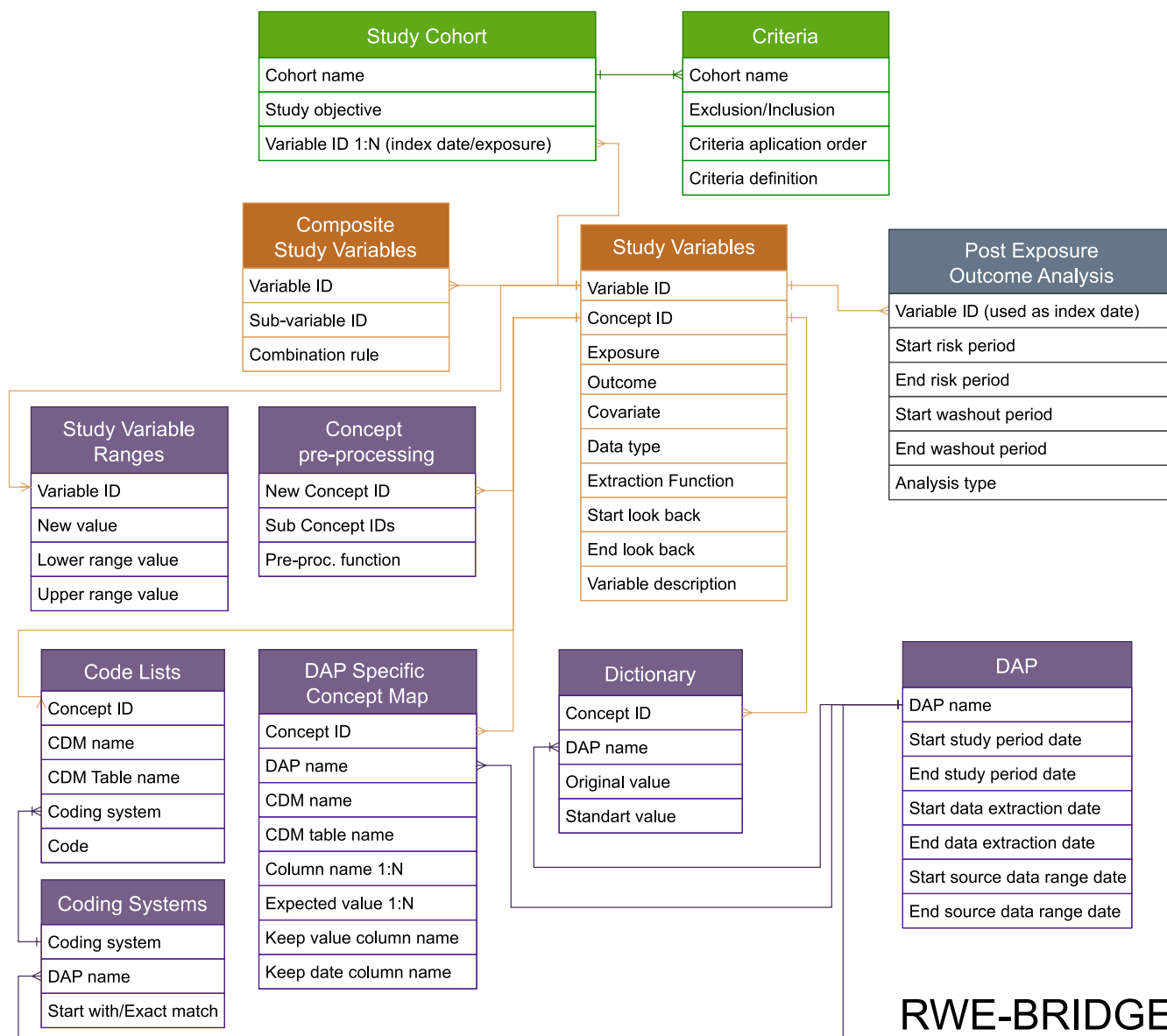
3.1.2.1 | DAP Specific Concept Map. This metadata table describes how to gather concepts without standardized codes that require combining multiple values stored in different columns or tables within the CDM instance. This table provides detailed information about each unstandardized concept in the participating DAP. The table also includes the name of the CDM table

where the concept records can be found, the name of the column in the CDM table where values are stored, and the columns to keep, usually values and date columns.

Following the example explained in 3.1.3. We can translate the content of the DAP Specific Map in a SQL query as:

```
SELECT person_id, mo_source_value, mo_date
FROM MEDICAL_OBSERVATIONS
WHERE mo_meaning='bmi' AND
      mo_unit='kg/m2'
```

3.1.2.2 | DAP. The DAP metadata table specifies the start and end dates for data extraction, source data range, and study period. This table allows the researcher to redefine base anchor dates if necessary for RWE studies [12]. As shown in



RWE-BRIDGE

FIGURE 3 | The RWE-BRIDGE diagram. The sections of the metadata schema are color-coded: Orange for the variable definition, green for the cohort definition, purple for the data retrieval and processing, and gray for the data analysis.

Figure 3, this table connects to all tables that contain information pertaining to DAPs; these are DAP specific concept map, dictionary, and coding systems.

3.1.2.3 | Coding Systems. The table provides information on the coding systems used by each DAP. The key field “Coding system” links it to the codelists table.

3.1.2.4 | Codelists. The Codelists are crucial for identifying various medical concepts, including symptoms, diagnoses, medicines, and more. They are linked to the Study Variables table via the “Concept ID” and contain vocabularies from coding systems like SNOMED CT, ICD10, and so on. They also specify the CDM table name and the actual codes. Comprehensive codelists allow programmers to select records from the CDM’s events table that match the codelist’s specific vocabularies using a simple SQL inner join:

```
SELECT EVENTS.person_id, CODE_LIST.concept,
EVENTS.code, EVENTS.date
FROM EVENTS
INNER JOIN CODE_LIST
WHERE EVENTS.code=CODE_LIST.code AND
EVENTS.coding_system=CODE_LIST.coding_system
```

If the codelist is incomplete—as opposed to the previous example—other approaches should be used to identify the code since every coding system might have different codification approaches (e.g., exact match, hierarchical).

3.1.2.5 | Concept Pre-Processing. The Concept pre-processing table creates or pre-processes new concepts by processing existing concepts from the Study Variables table. The New “Concept ID” key field links it to the Study Variables table. This is done by referencing programming functions to

the pre-processed concept. For example, the concept of BMI might need pre-processing if height and weight records are only available in the database. We can create a function that expects input weight and height and applies the formula.

$$\text{BMI} = \frac{\text{Weight(kg)}}{\text{Height(m)}^2}$$

resulting in a new concept. The pre-processing function is defined in the Pre-proc. function field, combined with the used concepts described in the Sub-concept ID field. When filling the pre-processing table, we add a new row per sub-concept, keeping the rest of the details the same.

Aligning with the KIP metric [20]. The pre-processed concepts are defined as having a KIP of (2, 2, 2). The KIP metric ranges from 0 (less pre-processed) to 2 (more pre-processed). In RWE-BRIDGE, a variable defined through a codelist has a KIP of (0, 0, 0), while those generated through a DAP Specific Concepts Map have a KIP of (1, 0, 0). Composite variables could be (-, 1, 1) or (-, 2, 1), with the knowledge conversion depending on the composing study variables.

3.1.2.6 | Dictionary. This table is used to standardize categorical values of study variables by replacing original data values with standardized ones. This is particularly useful when dealing with data from different DAPs that use different vocabulary. For instance, if DAP 1 uses “cm” and “m” to denote height, and DAP 2 uses “centimeters” and “meters”, the Dictionary table can be used to standardize these units to “cm” and “m” across all data. This standardization is crucial before performing calculations like BMI, which require consistent units of measurement. The Dictionary table is linked to the Study Variables table via the “Concept ID” field and to the DAP table via the “DAP Name” field.

3.1.2.7 | Study Variables Ranges. This post-processing table allows the categorizing of numerical variables into categories. The table defines the new value to be assigned and the range values for categorization. This table is used after the anchoring process when the values of the selected records are numerical and can be categorized. The reason why this step is happening after the anchoring and not before it is because there are numerical variables that result from a calculation within the anchoring window (i.e., the total number of medications between the index date and 30 days before), so the categorization needs to happen after anchoring.

3.1.3 | Cohort Definition

The section Cohort definition consists of two tables: Study cohort and criteria.

3.1.3.1 | Study Cohort. In this metadata table, the study subpopulations are described by providing names to the cohorts of interest, the study objective, and the start of the follow-up window for the population (t0). This table provides a method to define different populations and the observation and follow-up windows. Additionally, the study objective field allows for a free-text field to explain the use of the study cohort. This table

connects to the Study Variables table using the “Variable ID” field as the key.

3.1.3.2 | Criteria. This table outlines the in- and exclusion criteria, including the population to which each criterion applies, definitions, and application order. The Criteria connects to the Study Cohort table using the “Cohort name” field as the key. The criteria are expressions that can be used directly in the code. This way, the metadata table can be used—following the application order—to automate a data processing sequence for the study cohort.

3.1.4 | Data Analysis

The Data analysis and processing section consists of one table:

3.1.4.1 | Post-Exposure Outcome Analysis. This table outlines the various timeframes to consider when analyzing outcomes, such as calculating incidence or prevalence rates. The table includes the start and end dates for risk, control windows, and washout periods fields [13]. Risk windows are defined as the duration of treatment excluding the washout, whereas a washout period is a window of time between other periods; this prevents the undesired carry-over effects between periods of a patient [12]. The Post-exposure outcome analysis table comes with an “analysis type” column to provide further context on each study variable and it is linked to the Study Variables table through the key “Variable ID” field, from which the index date is used for the post-exposure analysis.

4 | Discussion

4.1 | Principal Findings

In this article, we describe the RWE-BRIDGE, a configuration tool that fills the methodological gap of the currently available tools/methodologies for code programming RWE studies. This tool builds upon previous initiatives ([6–9]), providing a ready-to-be-used, machine-readable and functional solution to the translation of study protocols into RWE analytical pipelines. (see Figure S1) In Figure S2, we illustrate how the information from the HARPER protocol (i.e., Tables) can be translated into the RWE-BRIDGE, facilitating the adoption of both tools. Thus, the RWE-BRIDGE improves the transparency of RWE studies’ analytical programming and communication with the researcher.

We have defined study variables and concepts following the KIP metric [20] as either simple, composite, or pre-processed, following an order based on the building complexity.

The DAP Specific Concept Map table facilitates semantic harmonization, retrieving unstandardized data (i.e., values) stored across multiple CDM tables within DAPs. Moreover, with the use of the Dictionary table and Concept pre-processing tables, these unstandardized values can be harmonized during the analytical programming. The Coding Systems, Codelist, and DAP tables provide an opportunity to cross-check the required information stated in SAP and the content on the DAP’s ETL data

instances; thus, the RWE-BRIGDE can also support the identification of upstream data issues before scripts are run locally.

4.2 | Strengths and Limitations

The RWE-BRIDGE structure is flexible, allowing new tables and fields per project while remaining machine-readable. An extra metadata table could list expected missing concepts per DAP, distinguishing between unavailable concepts (NA) and unidentified records (0) in our population.

The RWE-BRIDGE's CDM independence ensures CDM interoperability, meaning the metadata schema could be used in multi-database studies with one or more CDMs across DAPs. The DAP-specific variable Map table can prompt the settings for collecting the same variable from different CDMs. The example presented in Table 1 can be edited for the OMOP CDM (see Table 2). Furthermore, the Dictionary table allows for standardizing the categorical values across different CDMs.

The RWE-BRIDGE is compatible with any programming language, as this metadata schema can be structured in a set of different text files that can be loaded and edited by any programming language.

One limitation is that the manual population of the RWE-BRIDGE can be challenging, requiring knowledge of the SAP and metadata schema. While some fields in the metadata tables may appear unnecessary from a programmer's perspective, these ensure that each variable and analysis component is implemented as intended.

Another limitation of the RWE-BRIDGE is maintaining the consistency of the foreign keys across different tables during their population. To address this limitation, we developed the RWE-BRIDGE checker—details of the RWEBRIDGE-CHECKER can be found in the [Supporting Information](#).

The Post-Exposure Outcome Analysis table can be modified to fit other study designs better. The current metadata schema includes several tables designed to meet the requirements of post-authorization safety studies. However, it is possible to add or remove tables or fields from the schema to suit the needs of a particular study better—for example adding a table for descriptive analysis where each study variable is listed with its reporting categories.

4.3 | Implications for Researcher and Programmers

The RWE-BRIDGE is a reproducible configuration package designed specifically for RWE studies. Nonetheless, there is still room for improvement. The template is publicly available on GitHub (<https://github.com/UMC-Utrecht-RWE/RWE-BRIDGE>) and is open to comments.

The RWE-BRIDGE promotes transparency when implementing analytical scripts for RWE. This facilitates reproducibility and assures quality. For example, one can reuse phenotypes already

TABLE 1 | Example of DAP Specific Concept Map. In this example, we want to identify the body mass index (BMI) records for a specific DAP using the ConceptION CDM [18].

DAP name	CDM name	Concept ID	Table name	Field name 1	Field name 2	Field value 1	Field value 2	Field record date	Field record value
Example ConceptION	ConceptION	BMI	MEDICAL_OBSERVATIONS	mo_source_value	mo_unit	bmi	kg/m ²	mo date	mo value

Note: One can refer to the CDM table MEDICAL_OBSERVATIONS and ensure that the columns “medical observation meaning” (mo_meaning) and the column “medical observation unit” (mo_unit) have the values “BMI” and “m²”, respectively. This will identify each patient’s recorded BMI in the observation value column (mo_source_value) and date (mo_date).

TABLE 2 | Example of DAP Specific Concept Map for the OMOP CDM [21].

DAP name	CDM name	Concept ID	Table name	Field name 1	Field name 2	Field value 1	Field value 2	Field record date	Field record value
Example OMOP	OMOP	BMI	OBSERVATION	observation_concept_id	unit_source_column	bmi	kg/m ²	observation_date	value_source_value

available in phenotype libraries [22]. The RWE-BRIDGE can be made publicly available with a digital object identifier, adhering to the FAIR principles [23]. Furthermore, RWE-BRIDGE creates opportunities to apply standard procedures and create modular data engineering functionality following the recommendations of FAIR characteristics [24].

5 | Conclusions

Translating SAPs into analytical code is challenging. To overcome this problem, we developed RWE-BRIDGE to improve communication between researcher and programmers while promoting transparency and FAIRification of RWE.

Author Contributions

A.C.R., R.E., V.H., T.A.V., M.S., and C.L.A.N. conceived and designed the project. A.C.R., R.E., and V.H. carried out the investigation for the project development. Z.K. developed the metadata checker. T.A.V. and C.L.A.N. supervised. A.C.R. led the project and wrote the first draft of this manuscript. All authors revised the manuscript and approved the final version.

Acknowledgments

Thanks to Rosa Gini and the colleagues from RTI Health Solutions for the engaging discussions about programming challenges within the RWE field. We acknowledge the AstraZeneca (AZ) and AZ COVID-19 Vaccine PASS Research Teams for their essential contributions to the EUPAS43556 statistical analysis plan and the protocol referenced in this manuscript.

Conflicts of Interest

A.C.R., D.W., V.H., M.S., T.A.V., and C.L.A.N. are currently salaried employees at University Medical Center Utrecht, which receives institutional research funding from pharmaceutical companies and regulatory agencies and is administered by University Medical Center Utrecht. R.E. and Z.K. were salaried employees of University Medical Center Utrecht at the time this project was performed.

References

1. M. Li, S. Chen, Y. Lai, et al., "Integrating Real-World Evidence in the Regulatory Decision-Making Process: A Systematic Analysis of Experiences in the US, EU, and China Using a Logic Model," *Frontiers in Medicine* 8 (2021): 669509.
2. G. Trifirò, P. M. Coloma, P. R. Rijnbeek, et al., "Combining Multiple Healthcare Databases for Postmarketing Drug and Vaccine Safety Surveillance: Why and How?," *Journal of Internal Medicine* 275, no. 6 (2014): 551–561.
3. B. E. Maissenhaelter, A. L. Woolmore, and P. M. Schlag, "Real-World Evidence Research Based on Big Data: Motivation-Challenges-Success Factors," *Der Onkologe* 24, no. S2 (2018): 91–98.
4. S. T. Liaw, A. Rahimi, P. Ray, et al., "Towards an Ontology for Data Quality in Integrated Chronic Disease Management: A Realist Review of the Literature," *International Journal of Medical Informatics* 82, no. 1 (2013): 10–24.
5. R. Gini, M. C. J. Sturkenboom, J. Sultana, et al., "Different Strategies to Execute Multi-Database Studies for Medicines Surveillance in Real-World Setting: A Reflection on the European Model," *Clinical Pharmacology and Therapeutics* 108, no. 2 (2020): 228–235.
6. S. M. Langan, S. A. J. Schmidt, K. Wing, et al., "The Reporting of Studies Conducted Using Observational Routinely Collected Health

- Data Statement for Pharmacoepidemiology (RECORD-PE),” *BMJ* 363 (2018): k3532.
7. S. V. Wang, S. Pinheiro, W. Hua, et al., “STaRT-RWE: Structured Template for Planning and Reporting on the Implementation of Real World Evidence Studies,” *BMJ* 372 (2021): m4856.
8. N. M. Gatto, U. B. Campbell, E. Rubinstein, et al., “The Structured Process to Identify Fit-For-Purpose Data: A Data Feasibility Assessment Framework,” *Clinical Pharmacology and Therapeutics* 111, no. 1 (2022): 122–134.
9. S. V. Wang, A. Pottegård, W. Crown, et al., “HARmonized Protocol Template to Enhance Reproducibility of Hypothesis Evaluating Real-World Evidence Studies on Treatment Effects: A Good Practices Report of a Joint ISPE/ISPOR Task Force,” *Pharmacoepidemiology and Drug Safety* 32, no. 1 (2023): 44–55.
10. S. V. Wang, S. K. Sreedhara, S. Schneeweiss, and REPEAT Initiative, “Reproducibility of Real-World Evidence Studies Using Clinical Practice Data to Inform Regulatory and Coverage Decisions,” *Nature Communications* 13, no. 1 (2022): 5126.
11. N. M. Gatto, S. V. Wang, W. Murk, et al., “Visualizations Throughout Pharmacoepidemiology Study Planning, Implementation, and Reporting,” *Pharmacoepidemiology and Drug Safety* 31, no. 11 (2022): 1140–1152.
12. S. Schneeweiss, J. A. Rassen, J. S. Brown, et al., “Graphical Depiction of Longitudinal Study Designs in Health Care Databases,” *Annals of Internal Medicine* 170, no. 6 (2019): 398–406.
13. S. V. Wang, O. V. Patterson, J. J. Gagne, et al., “Transparent Reporting on Research Using Unstructured Electronic Health Record Data to Generate “Real World” Evidence of Comparative Effectiveness and Safety,” *Drug Safety* 42, no. 11 (2019): 1297–1309.
14. Center for Drug Evaluation and Research, “Best Practices for Conducting and Reporting Pharmacoepidemiologic Safety Studies Using Electronic Healthcare Data Sets,” U.S. Food and Drug Administration, FDA, 2020, <https://www.fda.gov/regulatory-information/search-fda-guidance-documents/best-practices-conducting-and-reporting-pharmacoepidemiologic-safety-studies-using-electronic>.
15. REPEAT, “REPEAT,” <https://www.repeatinitiative.org/>.
16. M. D. Wilkinson, M. Dumontier, I. J. J. Aalbersberg, et al., “The FAIR Guiding Principles for Scientific Data Management and Stewardship,” *Scientific Data* 3 (2016): 160018.
17. “Vaccine Monitoring Collaboration for Europe, VAC4EU,” <https://vac4eu.org/>.
18. N. H. Thurin, R. Pajouheshnia, G. Roberto, et al., “From Inception to ConcePTION: Genesis of a Network to Support Better Monitoring and Communication of Medication Safety During Pregnancy and Breast-feeding,” *Clinical Pharmacology and Therapeutics* 111, no. 1 (2022): 321–331.
19. A. Silberschatz, H. F. Korth, and S. Sudarshan, *Database System Concepts* (McGraw-Hill Education, 2011), <https://db-book.com/>.
20. N. Shang, C. Liu, L. V. Rasmussen, et al., “Making Work Visible for Electronic Phenotype Implementation: Lessons Learned From the eMERGE Network,” *Journal of Biomedical Informatics* 99 (2019): 103293.
21. P. E. Stang, P. B. Ryan, J. A. Racoosin, et al., “Advancing the Science for Active Surveillance: Rationale and Design for the Observational Medical Outcomes Partnership,” *Annals of Internal Medicine* 153, no. 9 (2010): 600–606.
22. M. Chapman, S. Mumtaz, L. V. Rasmussen, et al., “Desiderata for the Development of Next-Generation Electronic Health Record Phenotype Libraries,” *GigaScience* 10, no. 9 (2021): giab0599.
23. J. Weberpals and S. V. Wang, “The FAIRification of Research in Real-World Evidence: A Practical Introduction to Reproducible Analytic Workflows Using Git and R,” *Pharmacoepidemiology and Drug Safety* 33, no. 1 (2024): e5740.
24. D. Weibel, C. Dodd, O. Mahaux, et al., “ADVANCE System Testing: Can Safety Studies Be Conducted Using Electronic Healthcare Data? An Example Using Pertussis Vaccination,” *Vaccine* 38, no. S2 (2020): B38–B46.

Supporting Information

Additional supporting information can be found online in the Supporting Information section.